Optical Flow-based 3D Human Motion Estimation from Monocular Video

Thiemo Alldieck[1(✉)], Marc Kassubeck[1], Bastian Wandt[2], Bodo Rosenhahn[2], and Marcus Magnor[1]

[1]Computer Graphics Lab, TU Braunschweig
alldieck@cg.cs.tu-bs.de
[2]Institut für Informationsverarbeitung, Leibniz Universität Hannover

Abstract. This paper presents a method to estimate 3D human pose and body shape from monocular videos. While recent approaches infer the 3D pose from silhouettes and landmarks, we exploit properties of optical flow to temporally constrain the reconstructed motion. We estimate human motion by minimizing the difference between computed flow fields and the output of our novel flow renderer. By just using a single semi-automatic initialization step, we are able to reconstruct monocular sequences without joint annotation. Our test scenarios demonstrate that optical flow effectively regularizes the under-constrained problem of human shape and motion estimation from monocular video.

Fig. 1: Following our main idea we compute the optical flow between two consecutive frames and match it to an optical flow field estimated by our proposed optical flow renderer. From left to right: input frame, color-coded observed flow, estimated flow, resulting pose.

1 Introduction

Human pose estimation from video sequences has been an active field of research over the past decades with various applications such as surveillance, medical diagnostics or human-computer interfaces [22]. One branch of human pose estimation is referred

to as *articulated motion parsing* [41], which defines the combination of monocular pose estimation and motion tracking in uncontrolled environments. We present a new approach to temporally coherent human shape and motion estimation in uncontrolled monocular video sequences. Our work follows the *generative* strategy, where both pose and shape parameters of a 3D body model are found to match the input image through analysis-by-synthesis [21].

The 3D pose of a human figure is highly ambiguous when inferred from only a 2D image. Common generative approaches [14,15,8] try to find human poses that are a good match to given silhouettes. However, human silhouettes can often be explained by multiple poses [14]. Existing methods for landmark-based 3D human motion estimation from monocular images [25,33,1,39,40] can find a pose per frame independently. Although 3D reconstructions from both approaches look very convincing on single images, they can result in significant jumps in position and joint angles between two successive frames. This creates highly unrealistic 3D reconstructions in the temporal domain. Temporal consistency of tracked landmarks is only considered by few researchers [26,35,36].

In our work we exploit the properties of the optical flow in the sequence to not only enforce temporal coherence but also resolve the pose ambiguities of purely silhouette-based or landmark-based approaches. We develop a motion tracker based on our novel optical flow renderer. Optical flow has proven to improve 2D tracking while also sharing much of the properties of range data [29]. By exploiting properties of the optical flow we construct a robust and stable 3D human motion tracker working on monocular image sequences.

The main idea behind our work is that the optical flow between two consecutive frames largely depends on the change of the human pose between them. Following this idea, we propose an energy minimization problem that infers those model parameters that minimize the distance between observed and rendered flow for two input frames (Fig. 1). Additional energy terms are derived based on typical constraints of the human body, namely joint angle limits, limb interpenetration and continuous motion. For stable tracking, silhouette coverage is enforced.

We evaluate the proposed method using two well known datasets. We analyze the performance of our approach qualitatively and evaluate its 3D and 2D precision quantitatively. In the first experiment, 3D joint positions are compared against ground truth of the HumanEva-I dataset [32] and results of two recently published methods [5,36]. The second evaluation compares projected 2D joint positions against ground truth of the VideoPose 2.0 dataset [30] featuring camera movement and rapid gesticulation. We compare our results against a recent deep-learning-based method for joint localization [24]. Results demonstrate the strengths and potential of the proposed method.

Summarizing, our contributions are:

- We develop a novel optical flow renderer for analysis-by-synthesis.
- We propose a complete pipeline for 3D reconstruction of human poses from monocular image sequences, that is independent of previous annotations of joints. It only uses a single semi-automatic initialisation step.
- Optical flow is exploited to retrieve 3D information and achieve temporal coherence, instead of solely relying on silhouette information.

2 Related Work

Human pose estimation is a broad and active field of research. Here, we focus on model-based approaches and work that exploits optical flow during pose estimation.

Human pose from images. 3D human pose estimation is often based on the use of a body model. Human body representations exist in 2D and 3D. Many of the following methods utilize the 3D human body model SCAPE [2]. SCAPE is a deformable mesh model learned from body scans. Pose and shape of the model are parametrized by a set of body part rotations and low dimensional shape deformations. In recent work the SMPL model, a more accurate blend shape model compatible with existing rendering engines, has been presented by Loper et al. [20].

A variety of approaches to 3D pose estimation have been presented using various cues including shape from shading, silhouettes and edges. Due to the highly ill-posed and under-constrained nature of the problem these methods often require user interaction e.g. through manual annotation of body joints on the image. Guan et al. [14] have been the first to present a detailed method to recover human pose together with an accurate shape estimate from single images. Based on manual initialization, parameters of the SCAPE model are optimized exploiting edge overlap and shading. The work is based on [4], a method that recovers the 3D pose from silhouettes from 3-4 calibrated cameras. Similar methods requiring multi-view input have been presented, e.g. [3,31,27,10]. Hasler et al. [15] fit their own statistical body model [16] into monocular image silhouettes with the help of sparse annotations. Chen et al. [8] infer 3D poses based on learned shape priors. In recent work, Bogo et al. [5] present the first method to extract both pose and shape from a single image fully automatically. 2D joint locations are found using the CNN-based approach DeepCut [24], then projected joints of the SMPL model are fitted against the 2D locations. In contrast to our work no consistency with the image silhouette or temporal coherency is taken into consideration.

Pose reconstruction for image based rendering. 3D human pose estimation can serve as a preliminary step for image based rendering techniques. In early work Carranza et al. [7] have been the first to present free-viewpoint video using model-based reconstruction of human motion using the subject's silhouette in multiple camera views. Zhou et al. [38] and Jain et al. [18] present updates to model-based pose estimation for subsequent reshaping of humans in images and videos respectively. Rogge et al. [28] fit a 3D model for automatic cloth exchange in videos. All methods utilize various cues, none of them uses optical flow for motion estimation.

Optical flow based methods. Previous work has exploited optical flow for different purposes. Sapp et al. [30] and Fragkiadaki et al. [12] use optical flow for segmentation as a preliminary step for pose estimation. Both exploit the rigid structure revealing property of optical flow, rather than information about motion. Fablet and Black [11] use optical flow to learn motion models for automatic detection of human motion. Efros et al. [9] categorize human motion viewed from a distance by building an optical flow-based motion descriptor. Both methods label motion without revealing the underlying movement pattern. In recent work, Romero et al. [29] present a method for 2D human pose estimation using optical flow only. They detect body parts by porting the random forest approach used by the Microsoft Kinect to use optical flow. Brox et al. [6] have shown that optical flow can be used for 3D pose tracking of rigid objects. They propose

the use for objects *modeled as kinematic chains*. They argue that optical flow provides point correspondences inside the object contour which can help to identify a pose where silhouettes are ambiguous. Inspired by the above mentioned characteristics, we investigate the extent to which optical flow can be used for 3D human motion estimation from monocular video.

3 Method

Optical flow [13] is the perception of motion by our visual sense. For two successive video frames, it is described as a 2D vector field that matches a point in the first frame to the displaced point in the following frame [17]. Although calculated in the image plane, optical flow contains 3D information, as it can be interpreted as the projection of 3D scene flow [34]. Assuming the presence of optical flow in the sequence (i.e. all observed surfaces are diffuse, opaque and textured), the entire observed optical flow is caused by relative movement between object and camera. Besides the motion of individual body parts, optical flow contains information about boundaries of rigid structures and is an abstraction layer to the input images. Unique appearance effects such as texture and shading are removed [11,29]. We argue that these features make optical flow highly suitable for generative optimization problems.

The presented method estimates pose parameters (i.e. joint angles), global position, and rotation of a human model (Sec. 3.1) frame by frame. The procedure only requires a single semi-automatic initialization step (Sec. 3.6) and then runs automatically. The parameters for each frame are inferred by minimizing the difference between the observed and rendered flow (Sec. 3.3) from our flow renderer (Sec. 3.2). A set of energy functions based on pose constraints (Sec. 3.4) and silhouettes (Sec. 3.5) is defined to regularize the solution to meaningful poses and to make the method more robust.

3.1 Scene Model

In this work, we use the human body model SMPL [20]. The model can be reshaped using 10 shape parameters $\boldsymbol{\beta}$. For different poses, 72 pose parameters $\boldsymbol{\theta}$ can be set, including global orientation. $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ produce realistic vertex transformations and cover a large range of body shapes and poses. We define $(\gamma, \boldsymbol{\beta}, \boldsymbol{\theta}_i, \boldsymbol{\sigma}_i)$ as the model state at time step i, with global translation vector $\boldsymbol{\sigma}$ and gender γ. Here, for simplicity we assume that the camera positions and rotations as well as its focal lengths are known and static. It is however not required that the cameras of the actual scene are fixed, as the body model can rotate and move around the camera (cf. Sec. 4).

3.2 Flow Renderer

The core of the presented method is our differential flow renderer built upon OpenDR [19], a powerful open source framework for analysis-by-synthesis. The rendered flow image depends on the vertex locations determined by the virtual human model's pose parameters $\boldsymbol{\theta}$ and its translation $\boldsymbol{\sigma}$. To be able to render the flow *in situ*, we calculate the flow from frame i to $i-1$, referred to as backward flow. With this approach each

pixel, and more importantly, each vertex location contains the information where it came from rather than were it went and can be rendered in place. The calculation of the flow is achieved as follows: The first step calculates the displacement of all vertices between two frames i and j in the image plane. Then the flow per pixel is calculated through barycentric interpolation of the neighboring vertices. Visibility and barycentric coordinates are calculated through the standard OpenGL rendering pipeline.

The core feature of the utilized rendering framework OpenDR is the differentiability of the rendering pipeline. To benefit from that property, our renderer estimates the partial derivatives of each flow vector with respect to each projected vertex position.

3.3 Flow Matching

Having a flow renderer available, we can formulate the pose estimation as an optimization problem. The cost function E_f over all pixels p is defined as follows:

$$E_f = \sum_p ||F_{\mathrm{o}}(i, i-1, p) - F_{\mathrm{r}}(i, i-1, p)||^2 \tag{1}$$

where F_{r} refers to the *rendered* and F_{o} to the *observed* flow field calculated on the input frames i and $i-1$. The objective drives the optimization in such way that the rendered flow is similar to the observed flow (Fig. 1). As proposed in [19], we evaluate E_f not over the flow field but over its Gaussian pyramid in order to perform a more global search.

For this work we use the method by Xu et al. [37] to calculate the observed optical flow field. The method has its strength in the ability to calculate large displacements while at the same time preserving motion details and handling occlusions. The definition of the objective shows that the performance of the optical flow estimation is crucial to the overall performance of the presented method. To compensate for inaccuracies of the flow estimation and to lower the accumulated error over time, we do not rely exclusively on the flow for pose estimation, but employ additional constraints as well (Sec. 3.4 and Sec. 3.5).

3.4 Pose Constraints

SMPL does not define bounds for deformation. We introduce soft boundaries to constrain the joint angles in form of a cost function for pose estimation:

$$E_b = ||\max(e^{\boldsymbol{\theta}_{\min} - \boldsymbol{\theta}_i} - 1, 0) + \max(e^{\boldsymbol{\theta}_i - \boldsymbol{\theta}_{\max}} - 1, 0)||^2 \tag{2}$$

where $\boldsymbol{\theta}_{\min}$ and $\boldsymbol{\theta}_{\max}$ are empirical lower and upper boundaries and e and $\max$ are applied component-wise.

Furthermore, we introduce extended Kalman filtering per joint and linear Kalman filtering for translation. In addition to temporal smoothness, the Kalman filters are used to predict an *a priori* pose for the next frame before optimization, which significantly speeds up computation time.

During optimization the extremities of the model may intersect with other body parts. To prevent this, we integrate the interpenetration error term E_{sp} from [5]. The error term is defined over a capsule approximation of the body model. By using an error term interpenetration is not strictly prohibited but penalized.

Fig. 2: Method initialization. Observed image, manual pose initialization, first optimization based on joint positions (red: model joints; blue: manually marked joints), final result including silhouette coverage and optical flow based correction.

3.5 Silhouette Coverage

Pose estimation based on flow similarity requires that the rendered human model accurately covers the subject in the input image. Only body parts that cover the correct counterpart in the image can be moved correctly based on flow. To address inaccuracies caused by flow calculation, we introduce boundary matching.

We use the method presented by Bălan et al. [4] and adapt it to make it differentiable (cf. Sec. 3.7). A cost function measures how well the model fits the image silhouette S_I by penalizing non-overlapping pixels by the shortest distance to the model silhouette S_M. For this purpose Chamfer distance maps C_I for the image silhouette and C_M for the model are calculated. The cost function is defined as:

$$E_c = \sum_p ||aS_{M_i}(p)C_I(p) + (1-a)S_I(p)C_{M_i}(p)||^2 \tag{3}$$

where a weighs $S_{M_i}C_I$ stronger as image silhouettes are wider to enforce the model to reside within in the image silhouette than to completely cover it. To be able to compute derivatives, we approximate C_M by calculating the shortest distance of each pixel to the model capsule approximation used for E_{sp}. The distance at p is the shortest distance among all distances to each capsule. To lower computation time, we calculate only a grid of values and interpolate in between.

3.6 Initialization

For the initialization of the presented method two manual steps are required. First the user sets the joints of the body model to a pose that roughly matches the observed pose. It is sufficient that only the main joints such as shoulder, elbow, hip and knee are manipulated. In a second step the user marks joint locations of hips, knees, ankles, shoulders, elbows and wrists in the first frame. If the position of a joint cannot be seen or estimated it may be skipped. From this point no further user input is needed.

The initialization is then performed in three steps (Fig. 2). The first step minimizes the distance between the marked joints and their model counterparts projected to the image plane, while keeping E_{sp} and E_b low. We optimize over translation $\boldsymbol{\sigma}$, pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$. To guide the process we regularize both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ with objectives that penalize high differences to the manually set pose and the mean shape. In the second

step we include the silhouette coverage objective E_c. Finally, we optimize the estimated pose for temporal consistency. We initialize the second frame with the intermediate initialization result and optimize on the flow field afterwards. While optimizing E_f we still allow updates for $\boldsymbol{\theta}_0$ and $\boldsymbol{\sigma}_0$.

3.7 Optimization

After initialization we now iteratively find each pose using the defined objectives. The final objective function is a weighted sum of the energy terms of the previous sections:

$$\min_{\boldsymbol{\sigma},\boldsymbol{\theta}}(\lambda_f E_f + \lambda_c E_c + \lambda_b E_b + \lambda_{sp} E_{sp} + \lambda_M E_M) \tag{4}$$

with scalar weights λ. E_M regularizes the current state with respect to the last state

$$E_M = ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}||^2 + ||\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_{i-1}||^2. \tag{5}$$

Each frame is initialized with the Kalman prediction as described in Sec. 3.4.

For the optimization we use the OpenDR toolbox [19]. It allows for automatic differentiation of most partially differentiable functions. Therefore we can avoid the laborious and inaccurate task of calculating finite differences. All our energy terms are designed to be fully or partially differentiable. Using this auto-differentiation we are able to optimize Eq. (4) efficiently.

4 Evaluation

We evaluate the 3D and 2D pose accuracy of the presented method using two publicly available datasets: HumanEva-I [32] and VideoPose2.0 [30]. Ground truth is available for both datasets. We compare our results in both tests, 3D and 2D, against state-of-the-art methods [5,36,24]. Foreground masks needed for our method have been hand-annotated using an open-source tool for image annotation[1].

HumanEva-I. The HumanEva-I datasets features different actions performed by 4 subjects filmed under laboratory conditions. We reconstruct 130 frames of the sets *Walking C1* by subject 1 and *Jog C2* by subject 2 without reinitialization. The camera focal length is known. We do not adjust our method for the dataset except setting the λ weights. Fig. 3 shows a qualitative analysis. The green plots show the history of the joints used for evaluation. The traces demonstrate clearly the temporal coherence of the presented method. The low visual error in the last frames demonstrates that the presented method is robust over time.

We compare our method against the state-of-the-art methods of Bogo et al. [5] and Wandt et al. [36]. We use [5] without the linear pose regressor learned for the HumanEva sequences, which is missing in the publicly available source code. Frames that could not be reconstructed because of undetected joints have been excluded for evaluation. The 3D reconstruction of [36] is initialized with the same DeepCut [24] results as used for [5]. We measure the precision of the methods by calculating the *3D positioning error*

[1] https://bitbucket.org/aauvap/multimodal-pixel-annotator

Fig. 3: Resultant poses of frames 30 to 120 of the HumanEva-I test sets. Green traces show the history of evaluated joints.

as introduced by [33]. It calculates the mean euclidean distance of 13 reconstructed 3D joint locations to ground truth locations from MoCap data. Beforehand, optimal linear alignment of the results of all methods is achieved by Procrustes analysis. In order to demonstrate the global approach of our method, we follow two strategies here: First we measure the joint error after performing Procrustes alignment per frame. Afterwards we calculate a per sequence alignment over all joint locations in all frames and measure the resulting mean error. Table 1 shows the result of all tests.

The results show that our method performs best in three of four test scenarios. In contrast to [5] and [36], our method does not require prior knowledge about the performed motion or training of plausible poses. The better performance of our method can be explained by the temporal coherent formulation using optical flow. This strength is especially noticeable in the global analysis. The method of [5] takes no temporal consistency into consideration, which results in jumps of joint locations between two frames and unresolved pose ambiguities (cf. Fig. 4). Note that some frames cannot be reconstructed due to the joint detector failing to find a feasible skeleton. The algorithm of [36] also estimates the camera trajectory. A slightly wrongly estimated person size results in a global offset of the camera path and causes a larger global error. In order to demonstrate, that our method resolves ambiguities successfully, we conduct the experiment again with E_f set to zero. The resultant motion does no longer resemble the performed action (Fig. 5) and the positioning error raises significantly to 9.8 and 15.9 for local and global analysis of *Walking C1* and 14.5 and 22.3 for *Jog C2* respectively.

VideoPose2.0. After evaluation with fixed camera and under laboratory conditions, we test our method under a more challenging setting. The second evaluation consists of three clips of the VideoPose2.0 dataset. We choose the ”fullframe, every frame” (720×540px) variant in order to face camera movement. Ground truth is given in form of projected 2D location of shoulders, elbows, and wrists for every other frame. The camera focal length has been estimated.

We evaluate our method in 2D by comparison against DeepCut [24], the same method that has been used before as input for the 3D reconstruction methods. Table 2 shows the mean euclidean distance to ground truth 2D joint locations. We use the

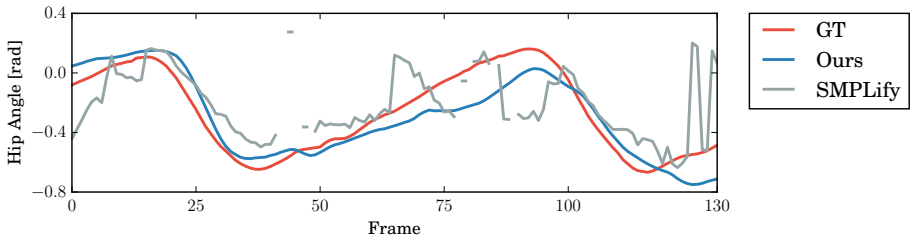

Fig. 4: Temporal behavior of the left hip angle of our method for *Walking C1* in comparison against ground truth (GT) and Bogo et al. (SMPLify) [5].

Fig. 5: Frame 120 of *Walking C1* in comparison to reconstruction with E_f set to zero.

first detected person by DeepCut and exclude several undetected joints from its evaluation. For our method, we project the reconstructed 3D joint locations to the image plane. The mixed performance of [24] is due to problems of the CNN with background objects. In order to enable fair comparison, we hand filter the results of [24] to foreground detections only and exclude several undetected joints. The comparison shows that our method produces similar precision while providing much more information. However, the increasing performance of CNN-based methods suggests that our method can benefit from semantic scene information for reinitialization in future work.

5 Conclusions

We have presented a new method for estimating 3D human motion from monocular video footage. The approach utilizes optical flow to recover human motion over time

Table 1: Mean 3D joint error in cm for local per frame Procrustes alignment and global per sequence alignment.

	Walking C1		Jog C2	
	local	global	local	global
Bogo et al. [5]	6.6	17.4	7.5	10.4
Wandt et al. [36]	5.7	34.0	**6.3**	38.0
Our method	**5.5**	**7.6**	7.9	**9.9**

Fig. 6: Resultant poses of frames 1, 21 and 41 of the VideoPose2.0 sets (Chandler, Ross, Rachel) with ground truth arm locations (green and blue).

Table 2: Mean 2D joint error (shoulders, elbows, and wrists) in pixels.

	Chandler	Ross	Rachel
DeepCut [24]	25.3	**10.5**	32.8
Our method	**23.3**	21.9	**15.9**

from a single initialization frame. For this purpose a novel flow renderer has been developed that enables direct interpretation of optical flow. The rich human body model SMPL provides the description of estimated human motion. Different test cases have shown applicability and robustness of the approach.

The presented method is dependent on realistic flow fields and good segmentation. It finds its natural limitations in the typical limits of optical flow estimation. Improvements in optical flow estimation, especially multi-frame optical flow, can help to further improve our method. Although our temporal coherent formulation allows for a good occlusion handling, large occlusions and reappearances can still lead to tracking errors.

Our work is focused on automatic estimation of human motion from monocular video. In future work we plan to further automatize our method. The method might benefit from recent developments in semantic segmentation [23] and joint angle priors [1]. Building upon the presented framework, the next steps are texturing of the model and geometry refinement, enabling new video editing and virtual reality applications.

Acknowledgments
The authors gratefully acknowledge funding by the German Science Foundation from project DFG MA2555/4-1.

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. IEEE Conference on Computer Vision and Pattern Recognition 2, 1446–1455 (2015)
2. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. In: ACM Transactions on Graphics (TOG). vol. 24, pp. 408–416. ACM (2005)
3. Bălan, A.O., Black, M.J., Haussecker, H., Sigal, L.: Shining a light on human pose: On shadows, shading and the estimation of pose and shape. In: IEEE International Conference on Computer Vision. pp. 1–8. IEEE (2007)
4. Bălan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision. Springer International Publishing (Oct 2016)
6. Brox, T., Rosenhahn, B., Cremers, D., Seidel, H.P.: High accuracy optical flow serves 3-d pose tracking: exploiting contour and flow based constraints. In: European Conference on Computer Vision. pp. 98–111. Springer (2006)
7. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.P.: Free-viewpoint video of human actors. In: ACM transactions on graphics (TOG). vol. 22, pp. 569–577. ACM (2003)
8. Chen, Y., Kim, T.K., Cipolla, R.: Inferring 3D shapes and deformations from single views. In: European Conference on Computer Vision. pp. 300–313. Springer (2010)
9. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision. pp. 726–733. IEEE (2003)
10. Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Marconiconvnet-based marker-less motion capture in outdoor and indoor scenes. IEEE transactions on pattern analysis and machine intelligence 39(3), 501–514 (2017)
11. Fablet, R., Black, M.J.: Automatic detection and tracking of human motion with a view-based representation. In: European Conference on Computer Vision. pp. 476–491. Springer (2002)
12. Fragkiadaki, K., Hu, H., Shi, J.: Pose from flow and flow from pose. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2059–2066 (2013)
13. Gibson, J.J.: The perception of the visual world. Houghton Mifflin (1950)
14. Guan, P., Weiss, A., Bălan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: International Conference on Computer Vision. pp. 1381–1388. IEEE (2009)
15. Hasler, N., Ackermann, H., Rosenhahn, B., Thormahlen, T., Seidel, H.P.: Multilinear pose and body shape estimation of dressed subjects from image sets. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1823–1830. IEEE (2010)
16. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. In: Computer Graphics Forum. vol. 28, pp. 337–346 (2009)
17. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence 17(1-3), 185–203 (1981)
18. Jain, A., Thormählen, T., Seidel, H.P., Theobalt, C.: MovieReshape: Tracking and reshaping of humans in videos. ACM Transactions on Graphics (TOG) 29(6), 148 (2010)
19. Loper, M.M., Black, M.J.: OpenDR: An approximate differentiable renderer. In: Computer Vision–ECCV 2014, pp. 154–169. Springer (2014)
20. Loper, M.M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics 34(6), 248:1–248:16 (Oct 2015)

21. Magnor, M.A., Grau, O., Sorkine-Hornung, O., Theobalt, C. (eds.): Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality. CRC Press (2015)
22. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding 104(2), 90–126 (2006)
23. Oliveira, G.L., Valada, A., Bollen, C., Burgard, W., Brox, T.: Deep learning for human part discovery in images. In: IEEE International Conference on Robotics and Automation (2016)
24. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2016)
25. Ramakrishna, V., Kanade, T., Sheikh, Y.A.: Reconstructing 3d human pose from 2d image landmarks. In: European Conference on Computer Vision (October 2012)
26. Rehan, A., Zaheer, A., Akhter, I., Saeed, A., Mahmood, B., Usmani, M., Khan, S.: Nrsfm using local rigidity. In: Winter Conference on Applications of Computer Vision. pp. 69–74. IEEE, Steamboat Springs, CO, USA (Mar 2014)
27. Rhodin, H., Robertini, N., Casas, D., Richardt, C., Seidel, H.P., Theobalt, C.: General automatic human shape and motion capture using volumetric contour cues. In: European Conference on Computer Vision. pp. 509–526. Springer (2016)
28. Rogge, L., Klose, F., Stengel, M., Eisemann, M., Magnor, M.: Garment Replacement in Monocular Video Sequences. ACM Trans. Graph. 34(1), 6:1–6:10 (2014)
29. Romero, J., Loper, M., Black, M.J.: Flowcap: 2D human pose from optical flow. In: Pattern Recognition, pp. 412–423. Springer (2015)
30. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1281–1288. IEEE (2011)
31. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: Advances in neural information processing systems. pp. 1337–1344 (2007)
32. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision 87(1-2), 4–27 (2010)
33. Simo-Serra, E., Ramisa, A., Aleny, G., Torras, C., Moreno-Noguer, F.: Single image 3d human pose estimation from noisy observations. In: Conference on Computer Vision and Pattern Recognition. pp. 2673–2680. IEEE (2012)
34. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. In: IEEE International Conference on Computer Vision. vol. 2, pp. 722–729. IEEE (1999)
35. Wandt, B., Ackermann, H., Rosenhahn, B.: 3D human motion capture from monocular image sequences. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (Jun 2015)
36. Wandt, B., Ackermann, H., Rosenhahn, B.: 3D reconstruction of human motion from monocular image sequences. Transactions on Pattern Analysis and Machine Intelligence (2016)
37. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(9), 1744–1757 (2012)
38. Zhou, S., Fu, H., Liu, L., Cohen-Or, D., Han, X.: Parametric reshaping of human bodies in images. In: ACM Transactions on Graphics (TOG). vol. 29, p. 126. ACM (2010)
39. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3d shape estimation from 2d landmarks: A convex relaxation approach. In: CVPR. pp. 4447–4455. IEEE Computer Society (2015)
40. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: Conference on Computer Vision and Pattern Recognition (June 2016)
41. Zuffi, S., Romero, J., Schmid, C., Black, M.J.: Estimating human pose with flowing puppets. IEEE International Conference on Computer Vision pp. 3312–3319 (2013)